MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

NCI Thesaurus: Using Science-Based Terminology to Integrate Cancer Research Results

Sherri de Coronado[a], Margaret W. Haber[b], Nicholas Sioutos[c], Mark S. Tuttle[d], Lawrence W. Wright[b]

[a]*Center for Bioinformatics, National Cancer Institute, Rockville, MD, USA*
[b]*Office of Communications, National Cancer Institute, Rockville, MD, USA*
[c]*Aspen Systems Corporation, Rockville, MD, USA*; [d]*Apelon, Inc., Alameda, CA, USA*

Abstract

Cancer researchers need to be able to organize and report their results in a way that others can find, build upon, and relate to the specific clinical conditions of individual patients. NCI Thesaurus ™ is a description logic terminology based on current science that helps individuals and software applications connect and organize the results of cancer research, e.g., by disease and underlying biology. Currently containing some 34,000 concepts – covering chemicals, drugs and other therapies, diseases, genes and gene products, anatomy, organisms, animal models, techniques, biologic processes, and administrative categories – NCI Thesaurus serves applications and the Web from a terminology server. As a scalable, formal terminology, the deployed Thesaurus, and associated applications and interfaces, are a model for some of the standards required for the NHII (National Health Information Infrastructure) and the Semantic Web.

Keywords:

Terminology, Cancer, Research, Information Systems, Description Logic, Terminology Server, Terminology Modeling.

Introduction

The last few decades have brought a dramatic increase in our understanding of cancer and related diseases at the biological, cellular and molecular levels, and cancer science is increasingly being shaped by molecular characterization of the biological processes related to cancer. The pace and complexity of this transformation requires new approaches to understand, manage, integrate and communicate current cancer-related concepts, so that results from basic and clinical research can better guide patient care as well as suggest new research directions. For example, the prognosis of certain myeloid leukemias and lymphomas can now be characterized by the unique chromosomal and molecular abnormalities they display, as in acute myeloid leukemia with t(8;21)(q22;q22), or ALK positive systemic anaplastic large cell lymphoma, in which the majority of cases have a t(2;5)(p23;q35) translocation involving the anaplastic large cell lymphoma kinase (ALK) gene. These and related "molecular" results have prompted investigators to revise both the naming and classification of many neoplasms. Ongoing research will result in the further cellular and sub-cellular identification of cohorts of patients with related clinical profiles. This explosion in understanding of the pathogenesis of cancer has resulted in great quantities of data and constant change of disease terms, creating confusion and complex management challenges. At least six classification schemes exist for non-Hodgkin's lymphoma, and a single type of lymphoma may be described using different terms from different classification schemes. [1,2]

To better organize, enhance and leverage the confusing patchwork of clinical and research terms in current use, the National Cancer Institute (NCI) has built the NCI Thesaurus™. Its goals are to: 1) provide a science-based terminology for cancer that is up-to-date, comprehensive, and reflective of the best current understanding; 2) make use of current terminology "best practices" to relate relevant concepts to one another in a formal structure, so that computers as well as humans can use the Thesaurus for a variety of purposes; and 3) speed the introduction of new concepts and new relationships in response to the emerging needs of basic researchers, clinical trials, information services and other users.

The NCI Thesaurus is part of the Enterprise Vocabulary Services (EVS) Project. [3] The EVS supports the standardization of vocabulary across the Institute and the larger, national and international cancer biomedical community. Befitting its wide role, formal governance and collaboration with current and prospective users of EVS services has major emphasis. At present, for example, a number of NCI and affiliated organizations help set priorities and define EVS content. In addition, EVS personnel are active in standards development organizations and interagency efforts to develop public domain vocabulary products and software in biomedicine and more generally. An early focus of such interagency efforts is the development of the NCI drug vocabulary in a way that will be re-used as part of a unified, open source federal drug information model [4] freely available for use by researchers, clinicians, and the public. NCI Thesaurus is an example of a potential standard with associated processes and applications that may play a role in the emerging NHII (National Health Information Infrastructure), [5] and it possesses some of the qualities of the ontologies and associated processes that will be required to support the Semantic Web. [6]

Materials and Methods

As part of its attempt to provide the best terminology content using the best available methods, NCI Thesaurus makes use of a combination of personnel, content, technology and processes not yet in wide-spread use. This combination is here broken out into

Management, Representation, Technology & Architecture, **Collaboration**, and **Process**.

Management

The Thesaurus project is led by NCI "cancer content" personnel with scientific and technical expertise, and most decisions are made by the appropriate members of a large inter-disciplinary team. As representatives of this team, the authors of this paper are trained in molecular biology, cancer care, pathology, computer science, and medical informatics. Frequently, multiple members of the project team are required to understand the causes of detected problems. If, for example, some terminology doesn't "look right" in an application, it's not always clear initially whether it's the terminology or the technology that is the proximal cause. At times the problem is the result of conflicts between different terminologies, and resolving the conflict requires deep domain knowledge.

Representation

Concept-based content for what became the NCI Thesaurus was initialized in 1997, first with a collection of local terminologies in use for coding documents related to managing science - funded grants, reports, and intramural science projects – and second with the PDQ Terminology[7][8] - used to code clinical trials, expert cancer summaries, and other resources used in NCI's public information services. The resulting vocabulary was placed into an NCI version of the UMLS Metathesaurus that supported synonymy, narrative definitions, mappings across standard vocabularies, and Web-based browsing. [9] Over time, as NCI terminology priorities focused more on basic and clinical research, other sources – such as ICD-O3 and current versions of MedDRA – were added. The NCI Metathesaurus environment continues to support a large intra- and extra-mural community.

To better enable consistent coding and retrieval across numerous, large NCI databases, and to help integrate the scientific results stored there, a more formally structured terminology – NCI Thesaurus – was required; hence, the subsequent move to a Description Logic-based terminology system.

Technology & Architecture

NCI Thesaurus is maintained and deployed using scalable, COTS (Commercial Off-The-Shelf) technology where appropriate, with public domain customizations as needed. Central to the COTS components are terminology editing and publishing tools from Apelon, including: TDE (Terminology Development Environment) with Workflow, and DTS (Distributed Terminology Server). [10] Analogously, the Apelon Metaphrase and Authority components are used to maintain the NCI Metathesaurus as shown in *Figure 1*. TDE Workflow enables periodic exports of change sets, conflict resolution, and publishing of new baselines. A new internal baseline is published weekly, and a new external baseline is published monthly. More recently, caBIO objects [11] in the caCORE 2.0 backbone [12] have been terminology enabled through a public domain API that provides full access to the basic and enhanced capabilities of both the NCI Thesaurus and NCI Metathesaurus servers.[13]

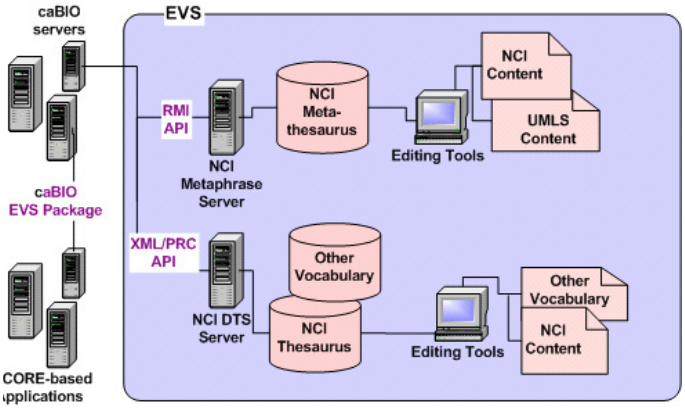

Figure 1 - NCI EVS Environment

The subset of DL manipulated by the TDE is defined by Apelon's Ontylog™ Data Model. The Data Model uses three basic components: *Concepts*, *Kinds* and *Roles*. *Concepts* are represented as nodes in an acyclic graph. *Roles* are directed edges defining relations between concepts. *Kinds* are disjoint sets of concepts used in role definitions to constrain domain and range. Each concept belongs to only one kind. For example, the role ***Disease_Has_Associated_Anatomy*** might have its domain restricted to the *Findings_**and_Disorders_Kind*** and its range to the ***Anatomy_Kind***.

All roles are passed from "superconcept" to concept in an inheritance hierarchy. For example, the concept "Breast Disorder" has the role ***Disease_Has_Associated_Anatomy*** connecting it to the concept "Breast". Since the concept "Breast Neoplasm" *Is_A* (superconcept) "Breast Disorder," "Breast Neoplasm" inherits the ***Disease_Has_Associated_Anatomy*** relation to the "Breast" concept, as do all the children of "Breast Neoplasm." These lateral (non-hierarchical) relations among concepts are referred to as associative or semantic roles – in contrast to (hierarchical) Is_A relationships.

Collaboration

Part of creating and deploying NCI Thesaurus in support of cancer care and research is making the best use of available resources. Toward this end, NCI exchanges drug, anatomy, and disease terminology, related technology, and best practices when possible and useful. Supporting these exchanges, NCI has in place informal working relationships or letters of understanding with other U.S. Federal Agencies such as VA (Department of Veterans Affairs), FDA (Food & Drug Administration), NLM (National Library of Medicine), and CDC (Centers for Disease Control). Initial collaboration has focused mainly on the assessment of shared terminology, both content and models, developed in common with other Federal agencies. Future collaboration will focus on how to most productively share the maintenance and enhancement burdens associated with this content. One early example of this kind of collaboration has been the development of extensions to common, shareable drug information in support of pharmacogenomics research. [14] Extension of such

shared terminology and models to support applications that interoperate between agencies is also being contemplated.

Process

Each of the previous **Method** elements – Management, Representation, Technology & Architecture, and Collaboration – can be viewed as NCI Thesaurus infra-structure that supports a *process*. To a good approximation, this process consists of iterations through the following cycle, with each cycle taking weeks to months: (1) refine the schema; (2) harvest existing terms, concepts, relationships and attributes; (3) add "Defining" and schema-specified relationships; (4) test NCI Thesaurus against users and applications; (5)identify content or schema shortfalls; and (6) repeat the process. Periodic creation and refinement of use cases – user scenarios – validates progress and sharpens priorities.

Results

The main NCI Thesaurus result is that all aspects of the Process are deployed and continuing on an ongoing basis; that is, NCI Thesaurus has rich and expanding formal, use-driven content, managed in a distributed environment by customized COTS tools, and "surfaced" by a COTS terminology server that supports a public domain API and Web Browser.

Schema

NCI Thesaurus content is organized as twenty *Is_A* hierarchies or kinds. A repertoire of approximately 50 types of role relationships provides the *differentiae* that define how concepts in the same kind are different from one another. Driven by user scenarios, these roles also facilitate translational research and support the bioinformatics infrastructure of the Institute. [15] Half of all concepts already have role values asserted or inherited.

Content

As shown in *Table 1*, the content of the NCI Thesaurus has grown rapidly as we have focused on areas of importance to our users. The content has been designed to satisfy Cimino's *Desiderata* [16], and it anticipates standards that will be required for the NHII (National Health Information Infrastructure) and the Semantic Web. Features include unique meaning-free concept identifiers, concepts modified or retired, but not deleted, and concept-based history tracking [17].

Table 1: NCI Thesaurus Content Statistics

Terms	Concepts	Category
1,691,850	783,158	NCI Metathesaurus
100,000	34,000	NCI Thesaurus
	10,521	Disease, Abnormality, Finding
	5,901	*Neoplasm (within Disease)*
	3,531	Drug
	2,773	Chemotherapy Regimen
	4,320	Anatomy
	1,767	Gene
	2,200	Protein (Gene Product)

NCI Thesaurus now provides extensive coverage of core interest areas for cancer researchers and other users, and provides a formal, use-driven organizing backbone for these topics. For example, types of cancer are being modeled using two main sets of "scientifically-defined" roles. First, the origin and context of the cancer are defined by five roles pointing to *Anatomy* values including primary and metastatic sites and the tissues and cells of origin. Second, the observed abnormalities that produce and express a particular neoplastic process are defined by roles pointing to the *Findings, Abnormal Cell, and Molecular Abnormality* hierarchies.

The following example - Mantle Cell Lymphoma - shows how these roles can characterize cancers. First are the biological and molecular features which define this particular cancer:

Anatomy: *Disease_Has_*	
_Primary_Anatomic_Site:	Lymphatic System
_Normal_Tissue_Origin:	Mantle zone
_Normal_Cell_Origin:	Mature B-Lymphocyte
Findings: *Disease_Has_*	
_Abnormal_Cell:	Centrocyte
_Cytogenetic_Abnormality:	t(11;14)(q13;q32)
_Molecular_Abnormality:	Monoclonal BCL-1 Gene Rearrangement
_Molecular_Abnormality:	Cyclin D1 mRNA Overexpression

Beyond this, a set of *Disease_May_Have…* roles cover important features which are often – but not always – true, in this case clinical findings of lymphadenopathy, hepatomegaly, and splenomegaly, and the cytogenetic abnormalities trisomy 12, del(13q14), and del(17p).

Figure 2, below, shows a schematic depiction of how the disease, drug, gene and protein kinds are being connected by roles to facilitate translational research, integrating data from varying domains and sources.

Uses and Users

As described above, NCI Thesaurus is one of two major vocabulary services provided by the NCI Enterprise Vocabulary Services (EVS), the other being the NCI Metathesaurus based on the NLM's UMLS Metathesaurus. NCI Metathesaurus contains numerous biomedical terminologies from the Metathesaurus, enhanced by the addition of the NCI Thesaurus, and specially licensed and local vocabularies needed by the NCI community. It provides the basis for mapping NCI cancer-related concepts to other vocabularies. NCI Thesaurus, on the other hand, is designed for database coding, search and data-mining, and is intended for developers of applications that need a tightly controlled vocabulary. The NCI Thesaurus is freely available for browsing [18], application access, and downloading. Public APIs, and distribution files of the Thesaurus as flat files and in Ontylog XML and OWL formats, are available through the NCI caCORE. These files are updated monthly, and include a concept history file.

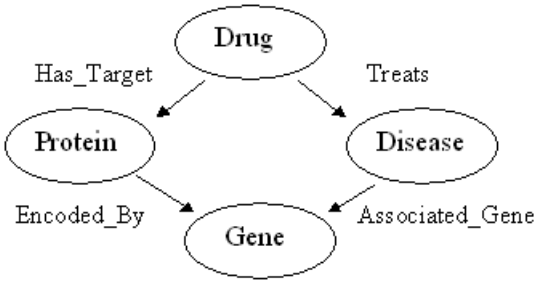

Figure 2 - Disease, Drug, Gene and Protein Modeling

NCI Thesaurus is used by a number of NCI applications to facilitate browsing, coding, and retrieval. Examples include *The Cancer Models Database* [19] and the cancer images portal caIMAGE [20]. These provide users with a tree-browser tool populated with the content of specifically chosen NCI Thesaurus sub-trees for browsing, coding new submissions, and retrieval. Tree traversal, explosion, and aggregation are enabled by the DL representation and can be manipulated to get a desired behavior from the content of the NCI Thesaurus. In these applications, the role relationships between anatomy and diseases enable coding and retrieval using either perspective. A user submitting a mouse model of prostate cancer can code a specific lobe of the prostate, but still be presented with all prostate diseases for selection. The data entry application checks to see if any diseases are associated with the selected anatomy – e.g.dorsal prostate gland, and, if not, the application walks back up the anatomy tree to "prostate gland" or until it finds a disease with the appropriate role relationship (*Disease Has Associated Anatomy*). Then it takes the highest level disease and all its children with that relationship (in this case "Prostate Disorder") to show the user. It is a simple use of the semantic relationships validating the utility of a DL terminology.

Current Areas of Content Development

First, because anatomy is a common, inter-disciplinary organizing principle, virtually every end user application needs simple, scientifically accurate, freely available anatomy terminology for annotating science result databases. We have recently enhanced and restructured the NCI Thesaurus anatomy, including a comprehensive microanatomy, to maximize our ability to relate this functionally organized domain with other related information. A partonomy browser has been developed to view both Part_of and Is_A, as well as other complex role relationships.

Second, we are modeling cancer-specific drugs, particularly those in trial, including the tracking of synonymous agent names and identifiers (from research codes to branded products) and providing definitions. Our goal is to pursue these and related activities within the emerging government framework for an open, shared drug terminology.

Third, we are working with the Mouse Models of Human Cancers Consortium (MMHCC) [21] to help meet terminology needs of those creating and using animal models. An early priority is enhancing diagnosis and, again, anatomy terminology with relevant links between mouse diseases and human diseases.

Fourth, relevant links are needed between cancer-related diseases and the genes or proteins involved in the disease process. Gene and protein databases abound, and the Gene Ontology Consortium provides terminology for biological process and molecular function that is useful for many purposes, but no other group is yet providing the specific links between diseases and genes and proteins related to cancer. Representing links between diseases and any related genes and proteins prompts the inevitable question regarding the boundaries between a terminology and a knowledge base. Our goal is to model only those "terminologic" things that are needed by humans or computers to specifically differentiate important concepts from one another, and those things that our researchers need to help them integrate and traverse data in their databases.

Discussion

NCI Thesaurus is an early example of scalable, "just-in-time" terminology. In only a few years it has gone from an idea to deployment, and current efforts are focused mainly on maturing the process used to maintain and enhance it in response to feedback from users. This rapid development was made possible by a mixture of "traditional" and novel terminology development and deployment practices. For example, by bootstrapping the initialization of NCI Thesaurus from existing terminologies, the project gained the co-operation of diverse stakeholders and avoided pitfalls associated with trying to develop a science-based terminology *de novo*. Similarly, by quickly surfacing early versions of Thesaurus in a terminology server, terminology developers and users, including application developers, could focus on "operational semantics" [22]. Users and developers are learning to think about NCI Thesaurus in terms of the behavior of the server in response to API calls. Thus, users and applications can deal with a "class" of tissue types or drugs while the precise enumeration of the class continues to evolve. Further, the history tracking mechanism of NCI Thesaurus can then tell users whether a concept they have coded has changed, and it can point them to the relevant merged, split, or retired concept.

As with most information technology artifacts, successful use increases demand for expansion of scope, for example, for a particular view of synonymy. The users may wish to "lump" concepts that are otherwise distinct – for example, to combine brand, generic, structural and functional drug class names into a single "drug" concept. Experiments are underway that should help us decide when such aggregations should be represented and maintained explicitly as data, and when they should be computed by applications. Not surprisingly, such decisions will affect NCI Thesaurus's maintenance burden and the ease with which its content can be shared with federal agencies and other stakeholders.

Conclusion

NCI Thesaurus has already become an important demonstration and testing ground for new approaches to terminology development and deployment. It is being integrated into numerous aspects of NCI's information services and systems, and provides standardized reference terminology for the biomedical data ob-

jects and metadata management built into NCI's caCORE cancer bioinformatics system. It is also being deployed and maintained at Web scale, becoming an important source of experience and collaboration with scalable, public domain, internationally accessible, human- and computer-enabling terminology.

Validation of the deepest hypothesis behind NCI Thesaurus – that a logic-based, science-driven Thesaurus can facilitate integration of cross-disciplinary information needed for cancer care and research – is underway. NCI Thesaurus is attempting to meet these needs in technology integration projects such as NCI's new Cancer Biomedical Informatics Grid (caBIG), [23] which is creating an infrastructure to integrate information technology and its use for major cancer research centers in the U.S. and beyond. Success in this sort of integration can make a real contribution to understanding, preventing, and treating cancer, and perhaps provide useful lessons for similar efforts in biomedicine.

Acknowledgements

The authors thank their EVS project colleagues, Frank Hartel, Gilberto Fragoso, Jim Oberthaler, Fred Rosenberg, Wen-Ling Shaiu, and other domain experts; John Carter and colleagues at Apelon; Ken Buetow and Peter Covitz at the NCI Center for Bioinformatics; and Gisele Sarosy and Richard Manrow of Cancer Information Products and Systems, NCI Office of Communications.

References

[1] Jaffe, ES, Harris NL, Stein H, Vardiman JW, eds. WHO Classification of Tumors of the Haematopoietic and Lymphoid Tissue. Lyon: IARC Press, 2001

[2] Harris, NL, Jaffe ES, Stein H, Banks PM, Chan JK, Cleary ML, Delsol G, Wolf-Peeters C, Falini B, Gatter KC. A revised European-American Classification of Lymphoid Neoplasms: A Proposal from the International Lymphoma Study Group. Blood 1994: 84:1361-92.

[3] NCI Enterprise Vocabulary Services (EVS). http://ncicb.nci.nih.gov/core/EVS

[4] Letter to Secretary HHS from Chair NCVHS, 2003 http://www.ncvhs.hhs.gov/031105lt2.pdf

[5] The National Health Information Infrastructure. http://aspe.hhs.gov/sp/nhii/ . See also W. A. Yasnoff, NHII: Moving Toward Implementation, http://www.med.utah.edu/medinfo/seminar.html

[6] W3C Technology and Society Domain, Semantic Web Activity. http://www.w3.org/2001/sw/.

[7] PDQ. http://cancer.gov/cancerinfo/pdq.

[8] Hubbard, SM, Martin NB, Thurn AL. NCI's Cancer Information Systems – Bringing Medical Knowledge to Clinicians. Oncology (Huntingt) 1995: 9(4):302-6.

[9] NCI Metathesaurus. http://ncimeta.nci.nih.gov.

[10]Apelon. http://www.apelon.com/products/products.htm

[11]http://ncicb.nci.nih.gov/core/caBIO .

[12]http://ncicb.nci.nih.gov/core

[13]CaCORE2.0 Technical Guide, ftp1.nci.nih.gov/pub/cacore/caCORE2.0_Tech_Guide.pdf

[14]Chute CG, Carter JS, Tuttle MS, Haber MW, Brown SH, Integrating Pharmacokinetics Knowledge into a Drug Ontology as an Extension to Support Pharmacogenomics. Proc AMIA Symp 2003.

[15]Covitz P, Hartel F, Schaefer C, de Coronado S, Fragoso H, Gustafson S, Buetow K. caCORE: A common infrastructure for cancer informatics. Bioinformatics 2003: 19(18):2404-2412.

[16]Cimino JJ. Desiderata for Controlled Medical Vocabularies in the Twenty-First Century, Meth Info Med 1998: 37, pp 394-403.

[17]Hartel F, Fragoso G, Ong K, Dionne R. Enhancing Quality of Retrieval Through Concept Edit History. Proc AMIA Symp 2003.

[18]NCI Thesaurus. http://nciterms.nci.nih.gov

[19]Cancer Models Database. http://cancermodels.nci.nih.gov/.

[20]caIMAGE Cancer Images Database. http://cancerimages.nci.nih.gov/caIMAGE/index.jsp

[21]MMHCC. http://emice.nci.nih.gov.

[22]Tuttle M, Campbell K, Keck K, Carter J. Toward Terminology as Infrastructure. Silva, J. ed. Cancer Informatics: Essential Technologies for Clinical Trials. NY: Springer-Verlag, 2002; pp.106-21

[23]caBIG. http://cabig.nci.nih.gov/.

Address for correspondence

Sherri de Coronado,

6116 Executive Blvd, #403, Bethesda, MD, 20892-8335